

Introduktion til Bioinformatik

Oversigt

Data & Databaser

Metoder

Opsamlende øvelse
Malaria vaccine

- Taksonomi
- DNA
- Protein
- Protein struktur
- Alignment
- Pairwise + Multiple
- BLAST (søgning)
- Fylogenetiske træer
- PyMOL (3D visualisering)

Kursusplan på vores wiki

Kursusforløb i Bioinformatik September 2011 - teaching

http://wiki.bio.dtu.dk/teaching/index.php/Kursusforløb_i_Bioinformatik_September_2011

Onsdag 21. September 2011

STED: Bygning 101 - lokale S02

11.30 - 12.00
Sandwich, kaffe mm.

12.00 - 12.15
Introduktion, præsentationsrunde af undervisere og kursister.

12.15 - 13.00
Foredrag: Introduktion – Evolution og DNA, biologisk information, DNA struktur og sekventering

- Baggrundsmateriale: "DNA Sequencing Tutorial" (PDF)
- Hand-out øvelse: "Base calling" (PDF).
- GenBank og FASTA fil format (PDF)
- Eukaryot gen-struktur (PDF).
- Slides: Introduktion, Evolution & DNA sekventering ([PowerPoint ↗])

Anders Gorm Pedersen ↗ (gorm@cbs.dtu.dk)

13.00 - 14.30
Øvelse: Søgning efter taksonomisk information i "Tree of Life" og "NCBI Taxonomy"

14.30 - 15.00
Demonstration: Søgning efter DNA sekvenser i GenBank databasen

Kaffepause (kaffe/the/vand + frugt/kage)

15.00 - 15.30
Foredrag: Proteiner og proteindatabaser

- Hand-out materiale: Proteinstrukturniveauer (PDF ↗)
- Slides: Proteinsekvenser & UniProt (Link kommer senere)

Anne Bresciani (agbr@bio.dtu.dk)

15.30 - 16.20
Øvelse: Translation af DNA sekvenser via Virtual Ribosome

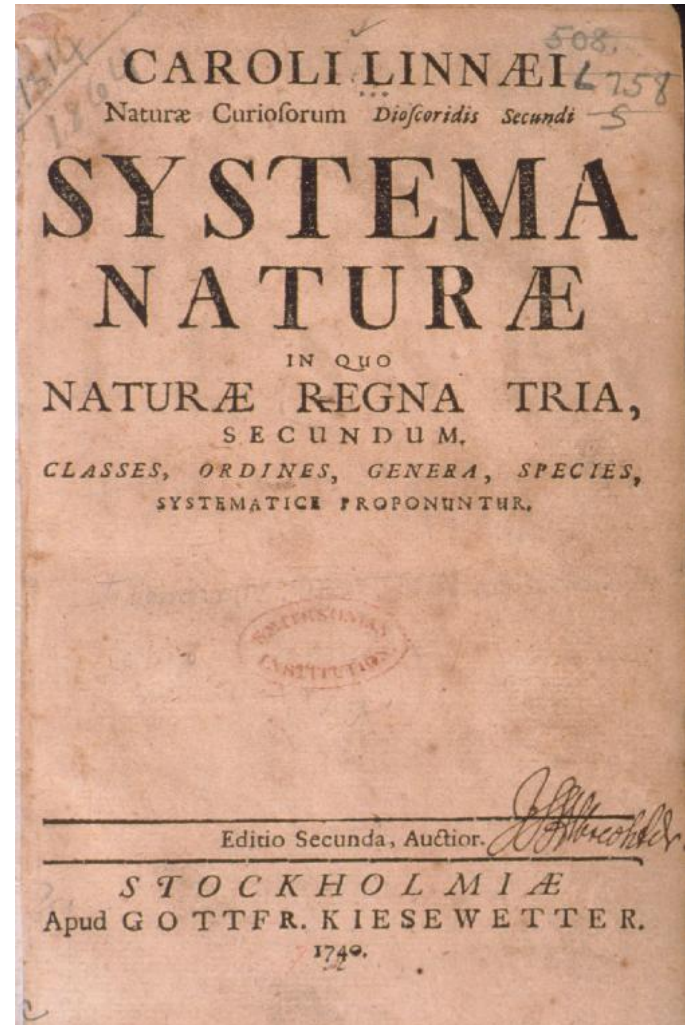
16.20 - 16.40
Hvordan kan bioinformatik implementeres i undervisningen i gymnasiet? ved Isa Kirk (isa@cbs.dtu.dk)

16.40 - 18.30
Øvelse: Søgning efter proteinsekvenser i UniProt databasen

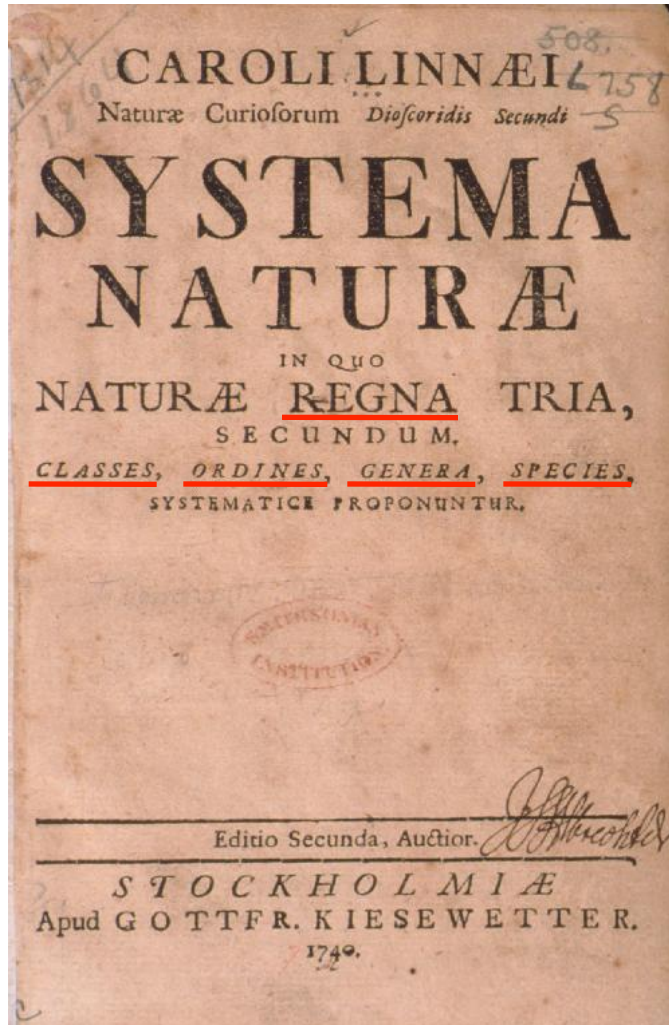
Classification: Linnaeus



Carl Linnaeus
1707-1778



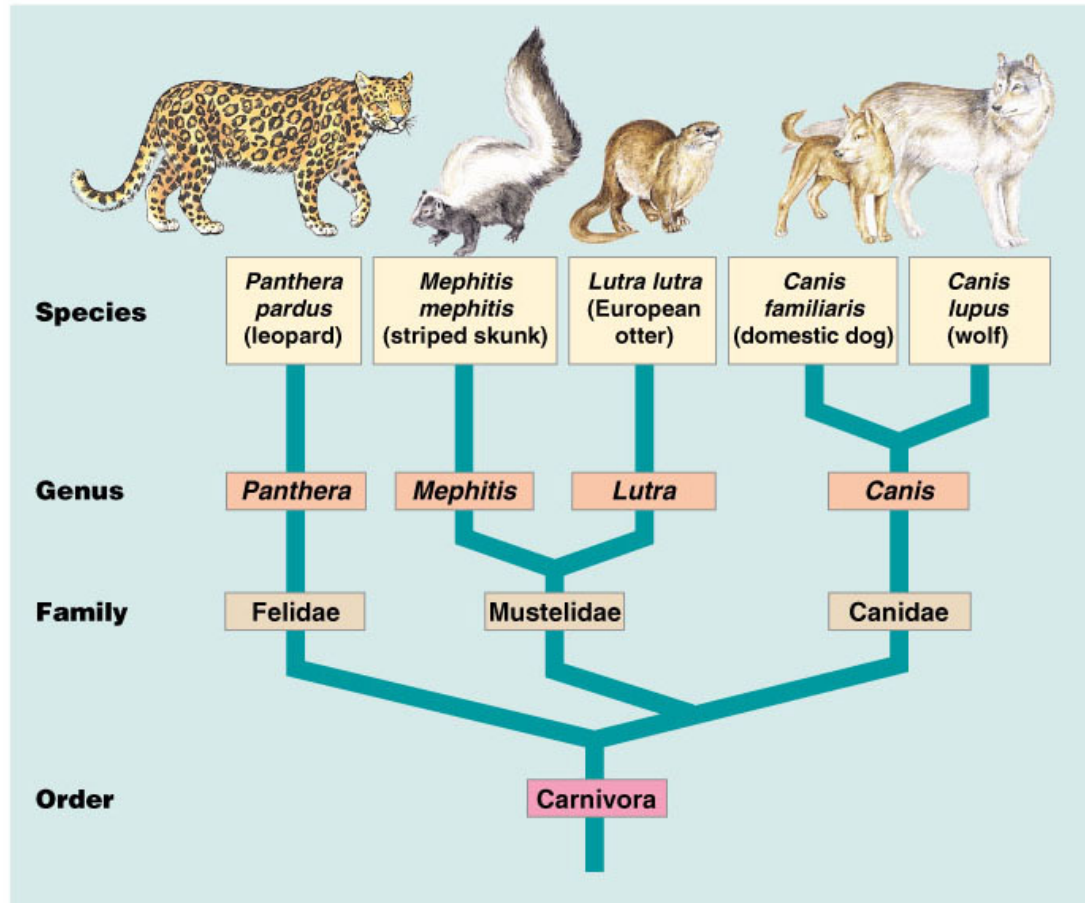
Classification: Linnaeus



Hierarchical system

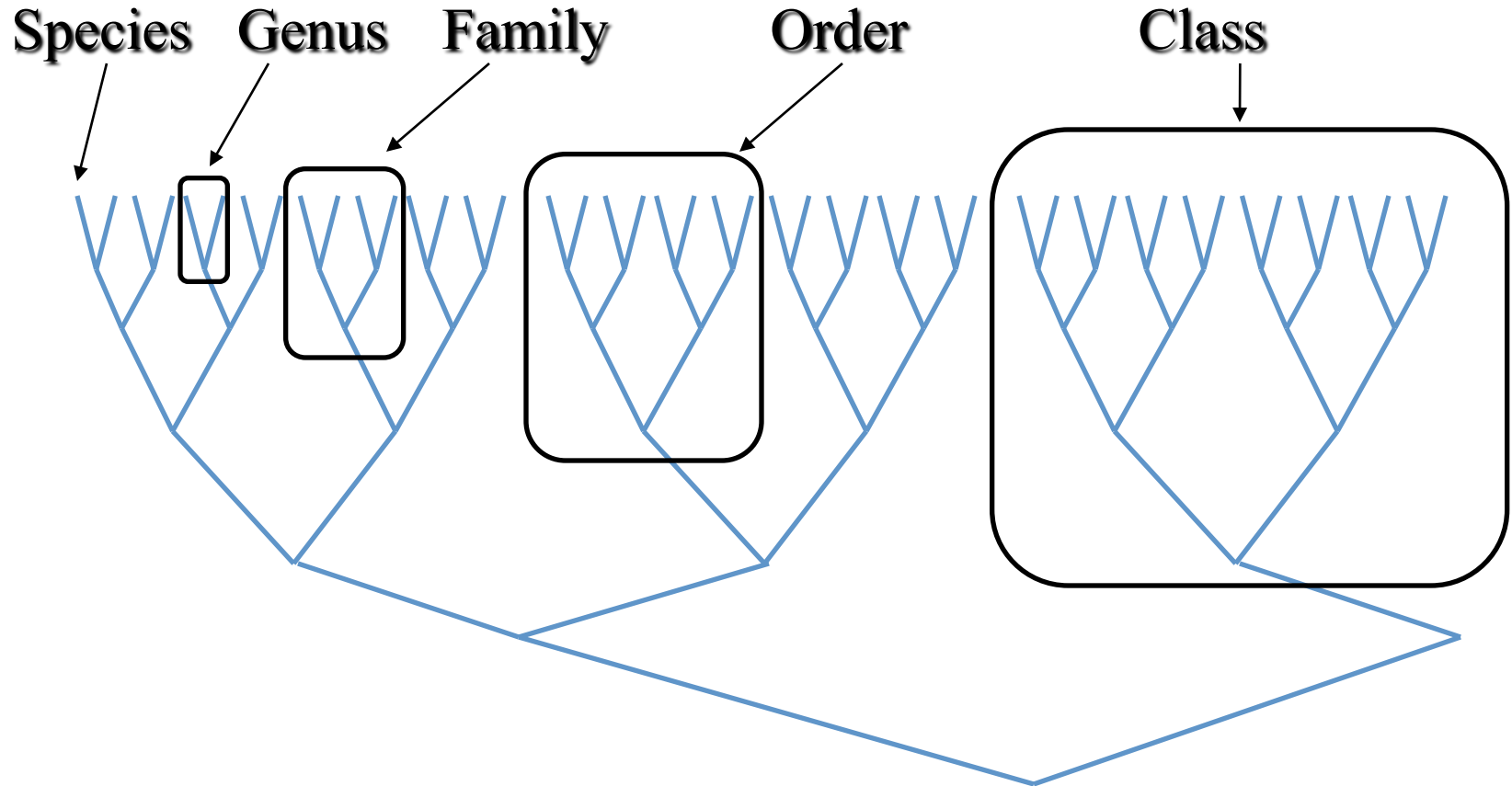
- Kingdom (Rige)
- Phylum (Række)
- Class (Klasse)
- Order (Orden)
- Family (Familie)
- Genus (Slægt)
- Species (Art)

Classification depicted as a tree

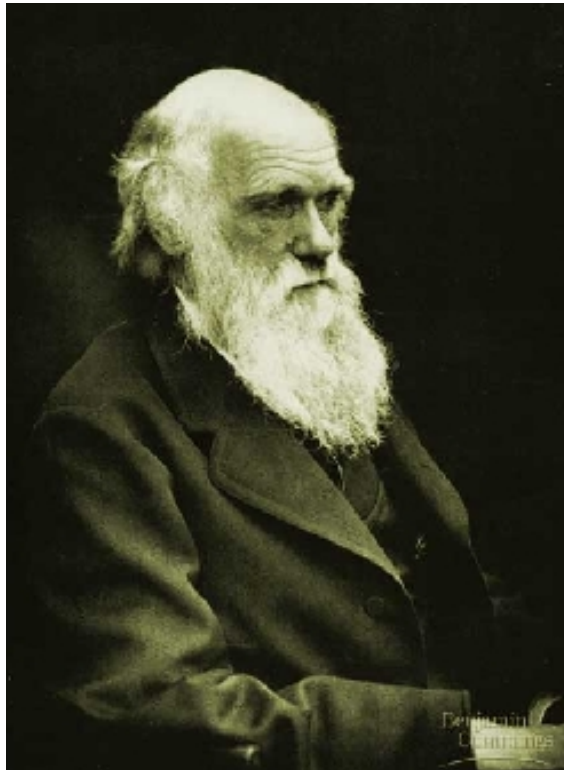


Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

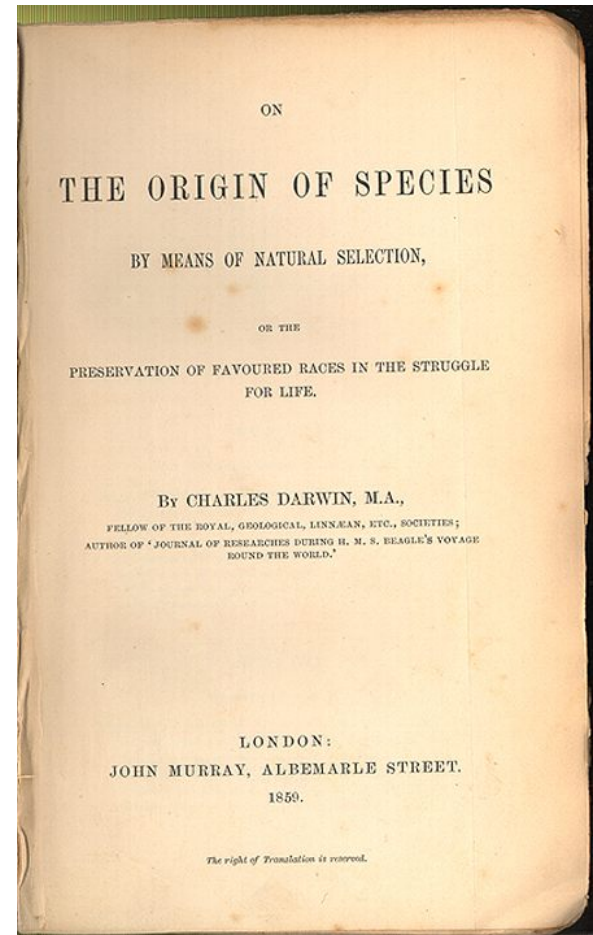
Classification depicted as a tree



Theory of evolution

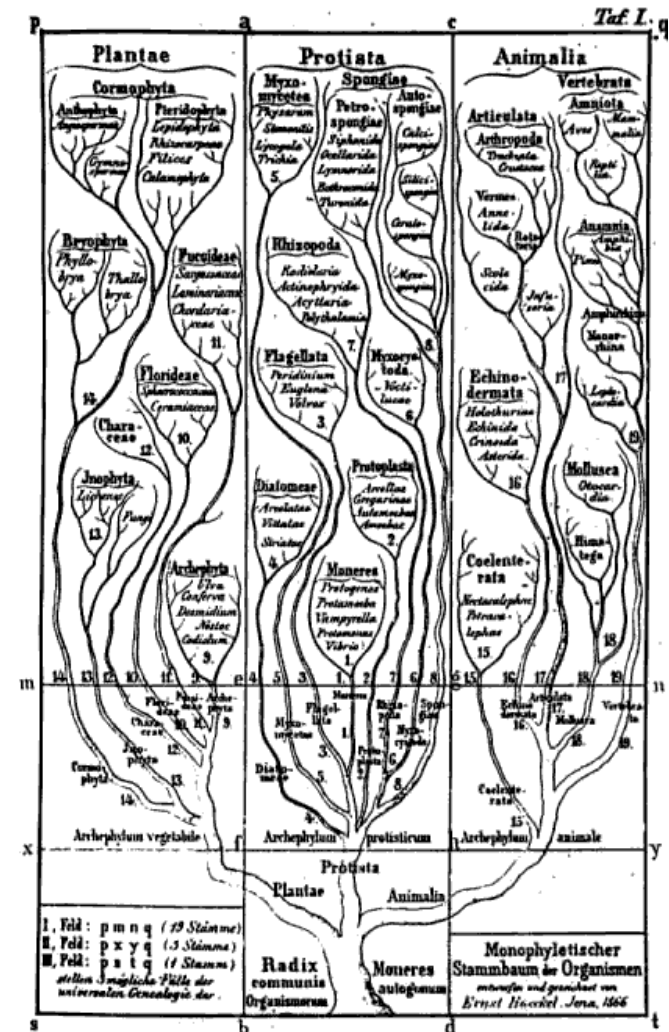


Charles Darwin
1809-1882



Phylogenetic basis of systematics

- Linnaeus:
Ordering principle is God.
- Darwin:
Ordering principle is shared descent from common ancestors.
- Today, systematics is explicitly based on phylogeny.



Natural Selection: Darwin's four postulates

More young are produced each generation than can survive to reproduce.

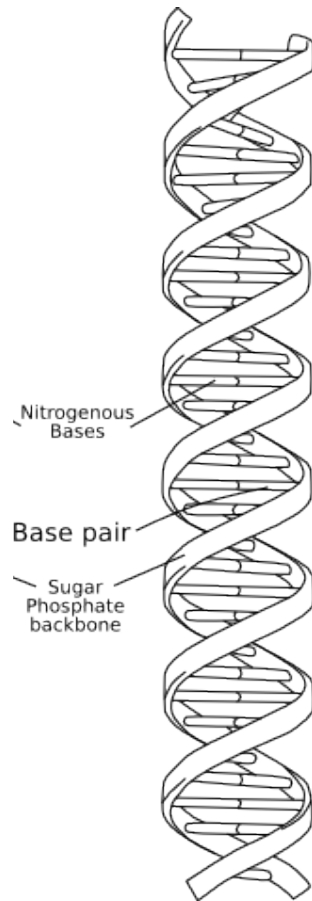
Individuals in a population vary in their characteristics.

Some differences among individuals are based on genetic differences.

Individuals with favorable characteristics have higher rates of survival and reproduction.

- Evolution by means of natural selection
- Presence of "design-like" features in organisms:
- Quite often features are there "for a reason"

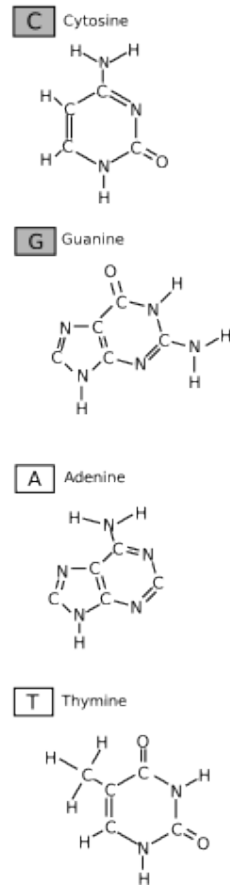
Molecular Basis for Heredity: DNA



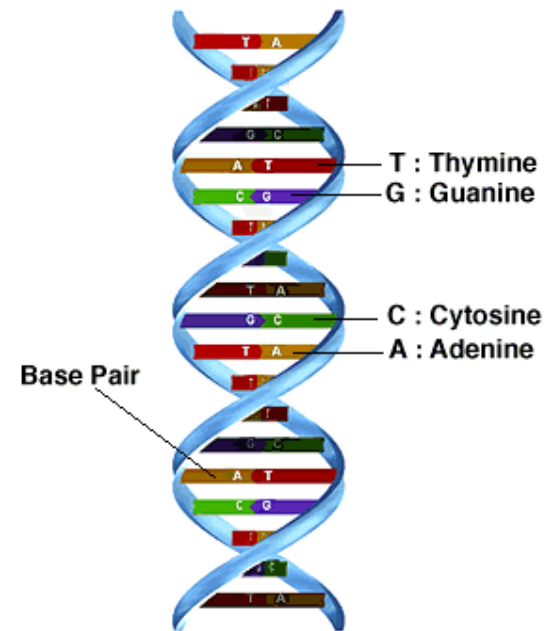
DNA

d

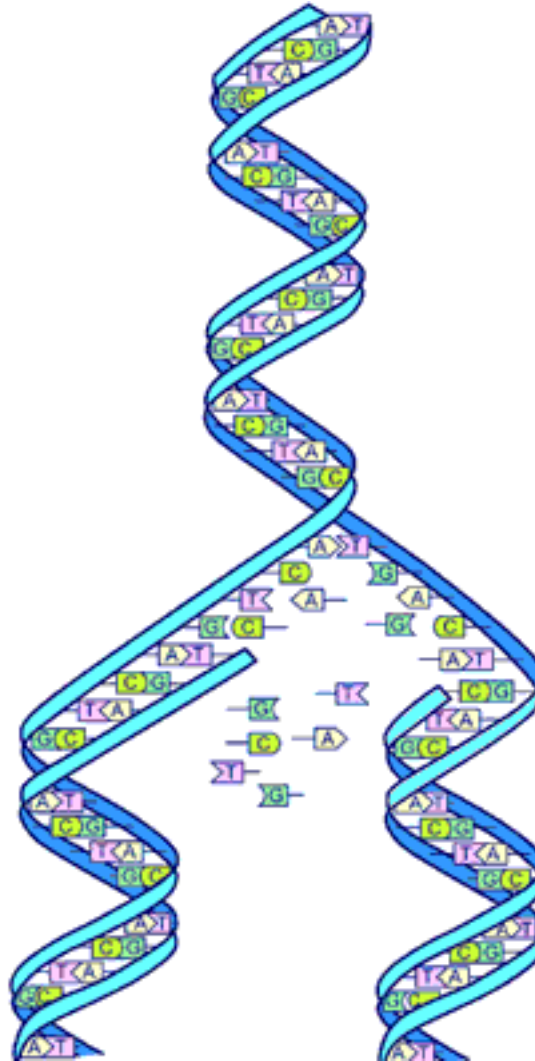
Deoxyribonucleic acid



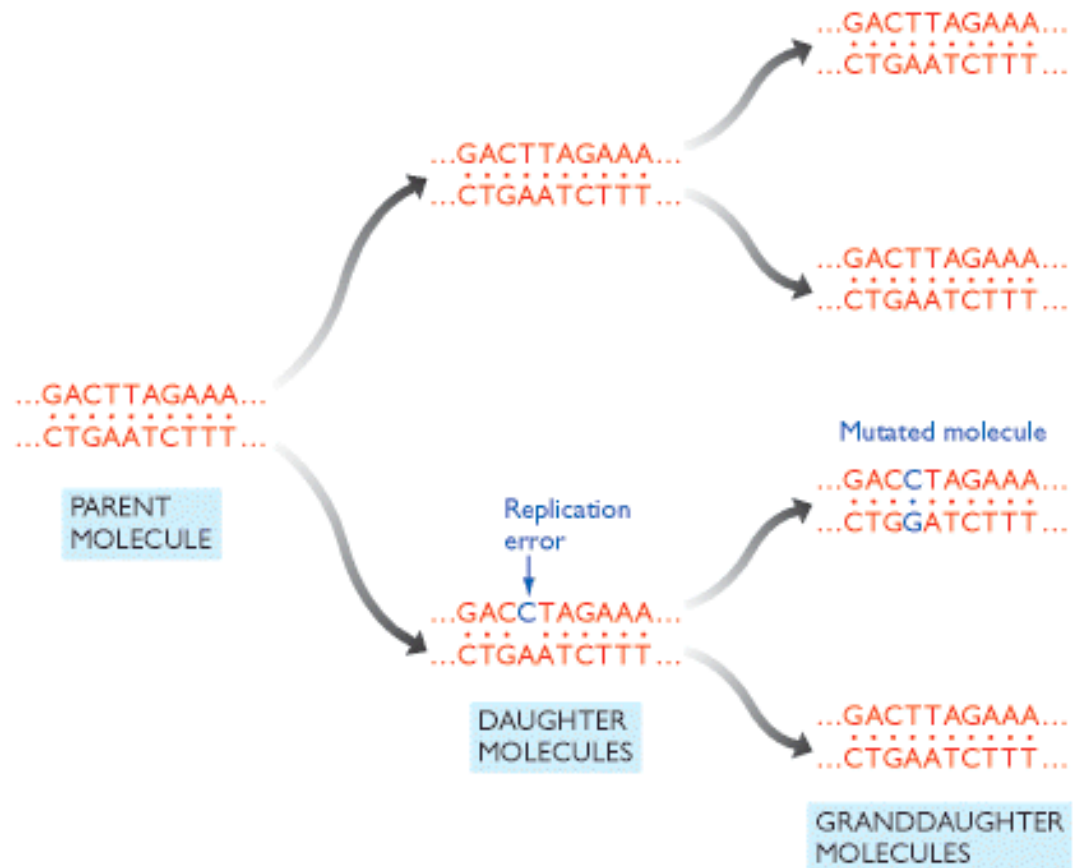
Nitrogenous Bases



Molecular Basis for Heredity: DNA



Molecular Basis for Variation: DNA Mutation



A history of mutations

ATGGCAATGTG**G**ATGCA

ATGGCCCC**C**GTG**G**AACCG

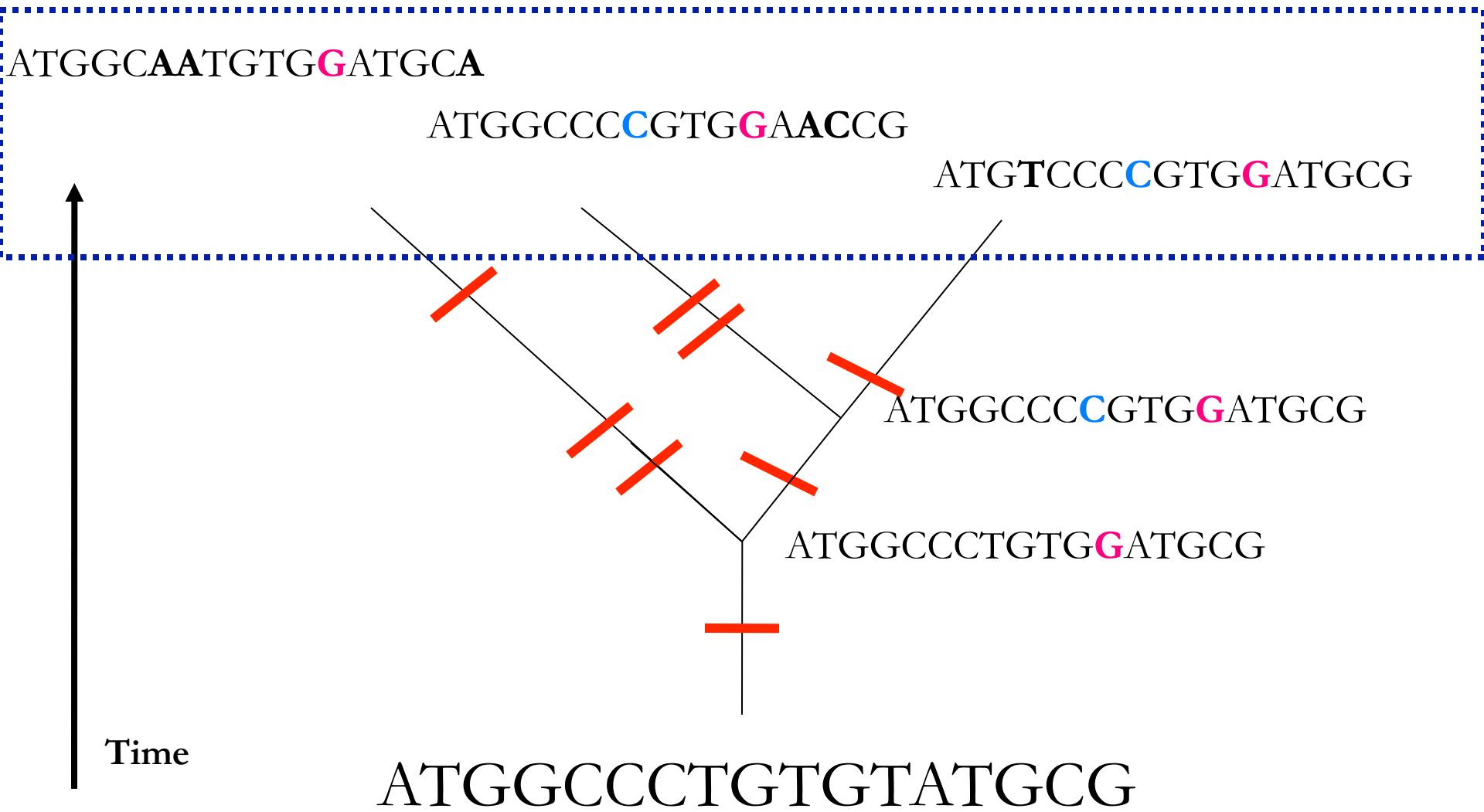
ATGTCCCC**C**GTG**G**ATGCG

ATGGCCCC**C**GTG**G**ATGCG

ATGGCCCTGTG**G**ATGCG

ATGGCCCTGTGTATGCG

Time



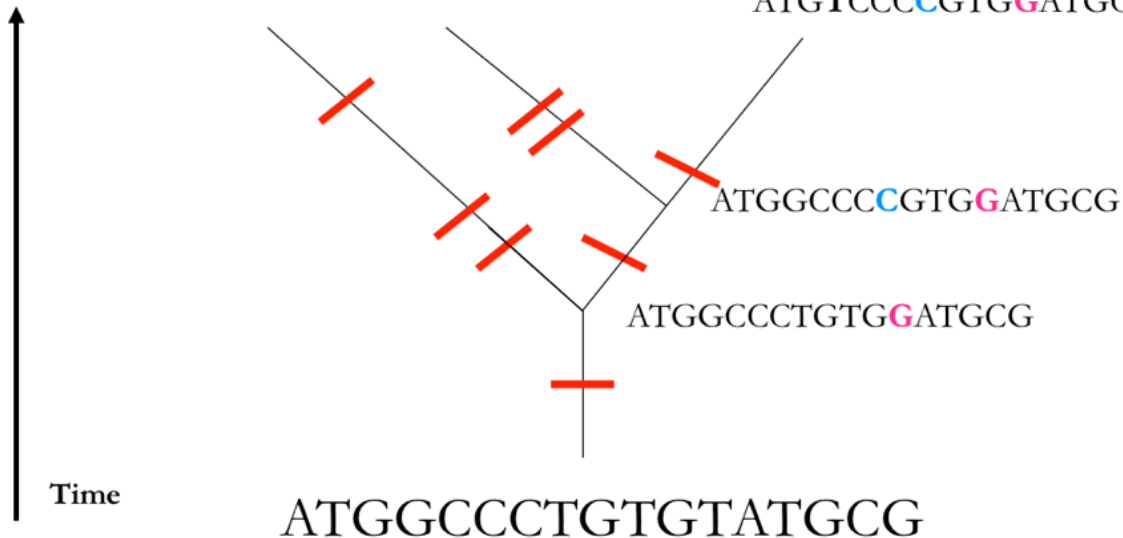
“DNA alignment”

- Species1: ATGGCAATGTG**G**ATGCA
 - Species2: ATGGCCCC**C**GTG**G**AACCG
 - Species3: ATGTCCCC**C**GTG**G**ATGCG
- $\left. \begin{array}{l} 6 \\ 3 \end{array} \right\} 5$

ATGGCAATGTG**G**ATGCA

ATGGCCCC**C**GTG**G**AACCG

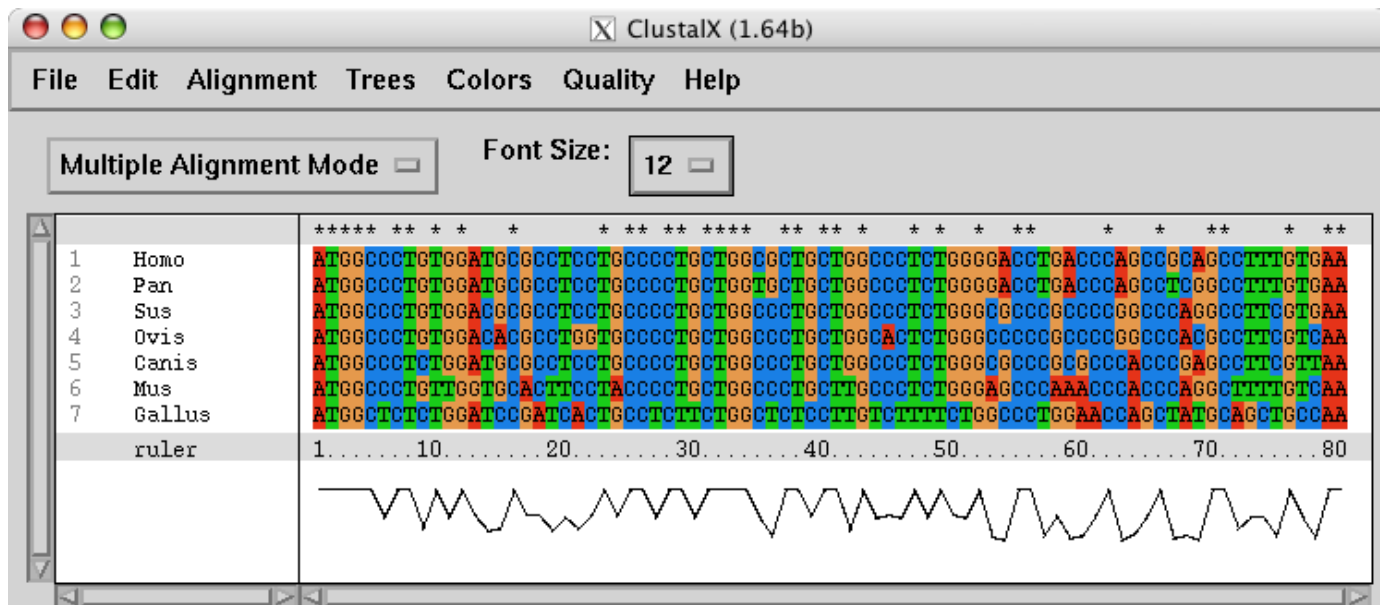
ATGTCCCC**C**GTG**G**ATGCG

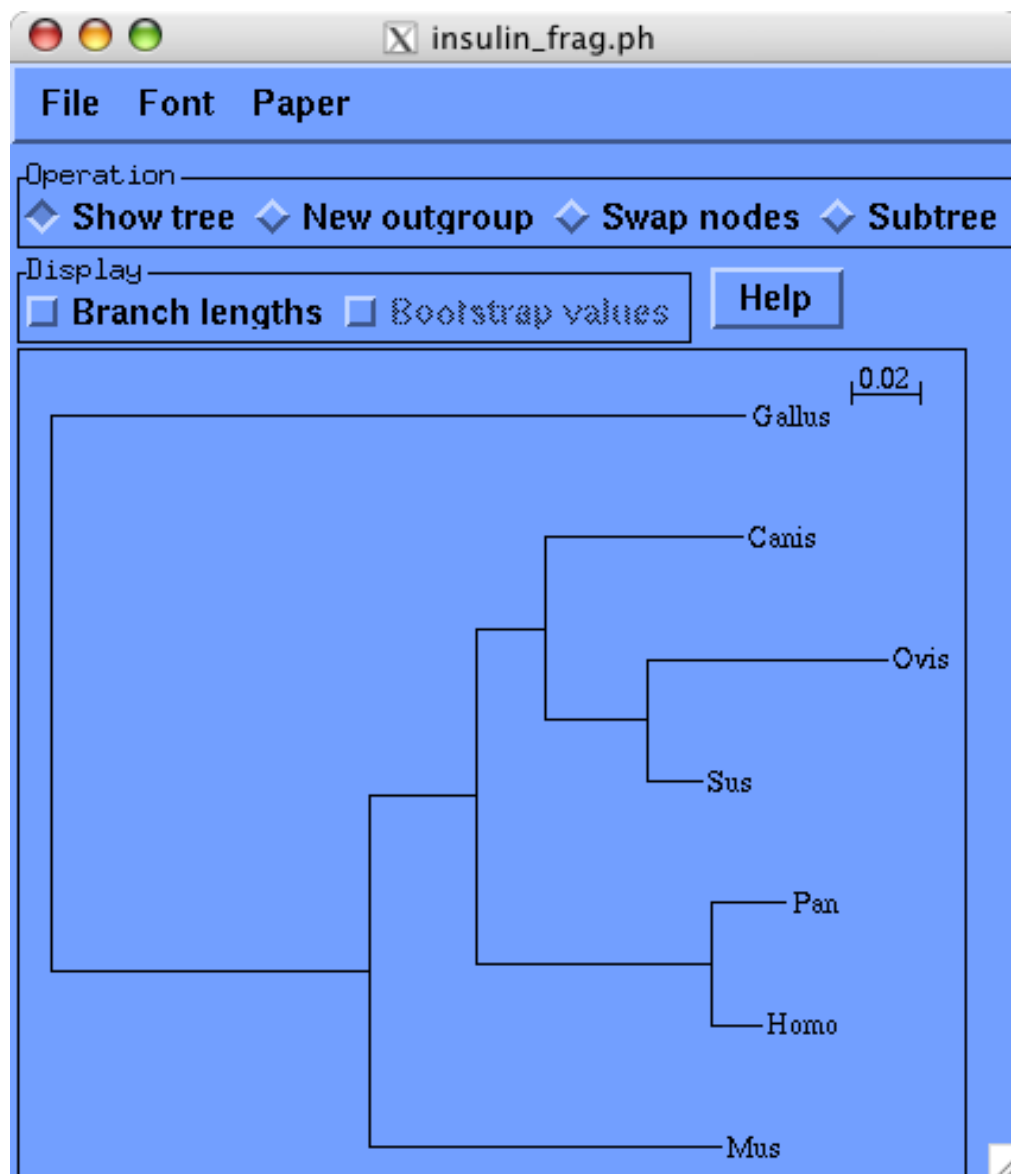


Real life example: Alignment

- Insulin from 7 different species

- Homo: ATGGCCCTGTGGATGCGCCTCCTGCCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAA
- Pan: ATGGCCCTGTGGATGCGCCTCCTGCCCCCTGCTGGTGTCTGCTGGCCCTCTGGGGACCTGACCCAGCCTCGGCCTTTGTGAA
- Sus: ATGGCCCTGTGGACGCGCCTCCTGCCCCCTGCTGGCCCTGCTGGCCCTCTGGGCGCCCCGCCCGGCCAGGCCTTCGTGAA
- Ovis: ATGGCCCTGTGGACACGCCTGGTGGCCCTGCTGGCCCTGCTGGCACTCTGGGCCCCCGCCCCGGCCCACGCCTTCGTCAA
- Canis: ATGGCCCTCTGGATGCGCCTCCTGCCCCCTGCTGGCCCTGCTGGCCCTCTGGGCGCCCCGCGCCCACCCGAGCCTTCGTAA
- Mus: ATGGCCCTGTTGGTGCACCTCCTACCCCTGCTGGCCCTGCTTGCCTCTGGGAGCCCAAACCCACCCAGGCCTTTGTCAA
- Gallus: ATGGCTCTCTGGATCCGATCACTGCCTCTTCTGGCTCTCCTTGTCTTTTCTGGCCCTGGAACCAGCTATGCAGCTGCCAA

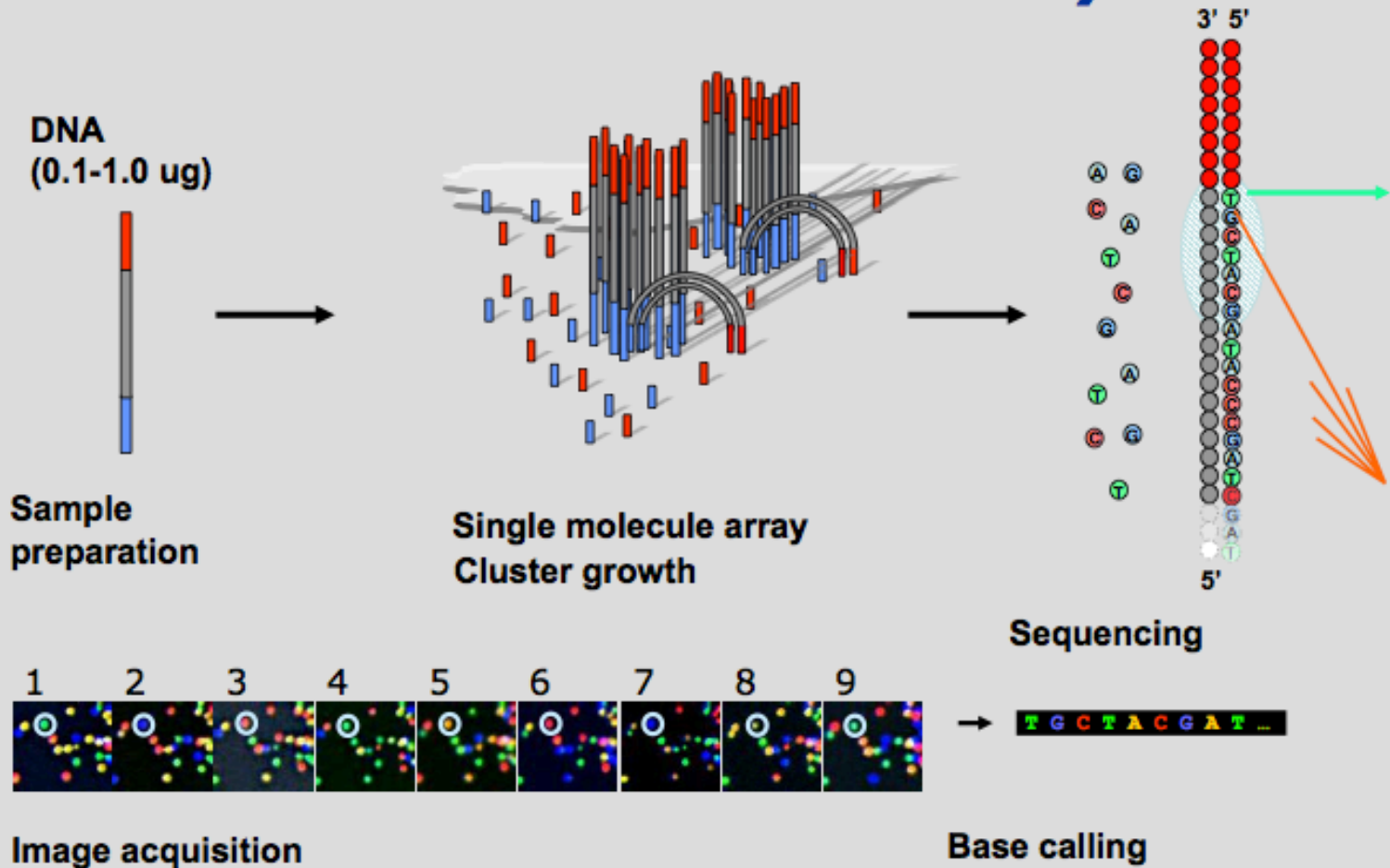




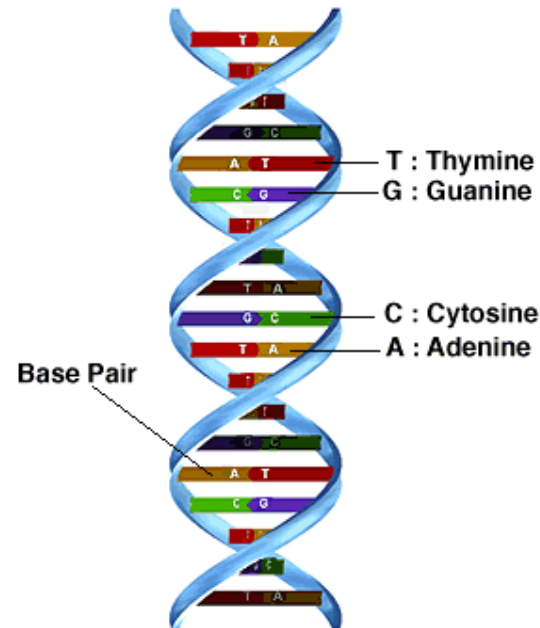
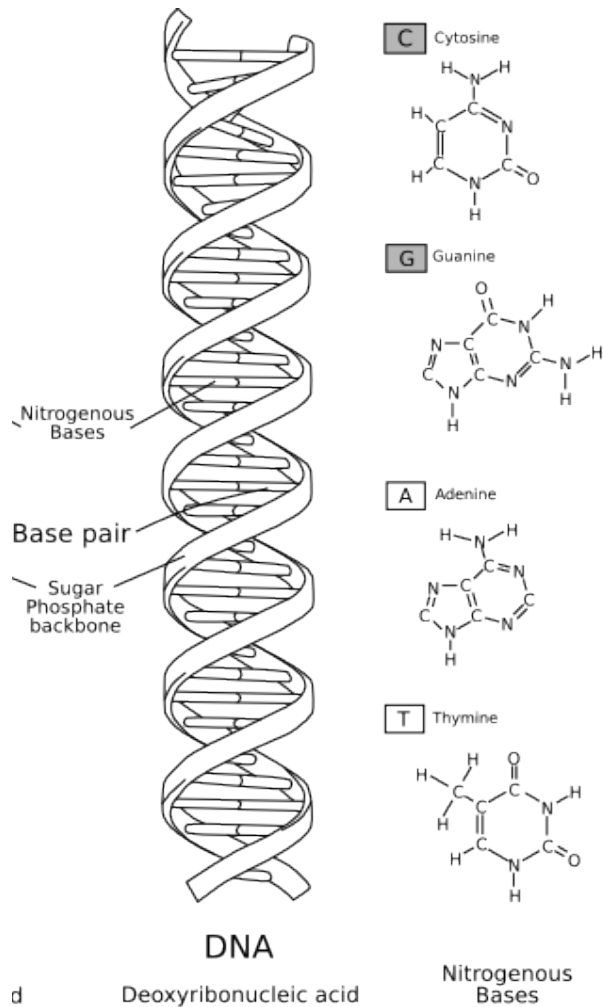


Illumina Sequencing Technology

Reversible Terminator Chemistry



Symbolic representation of DNA structure



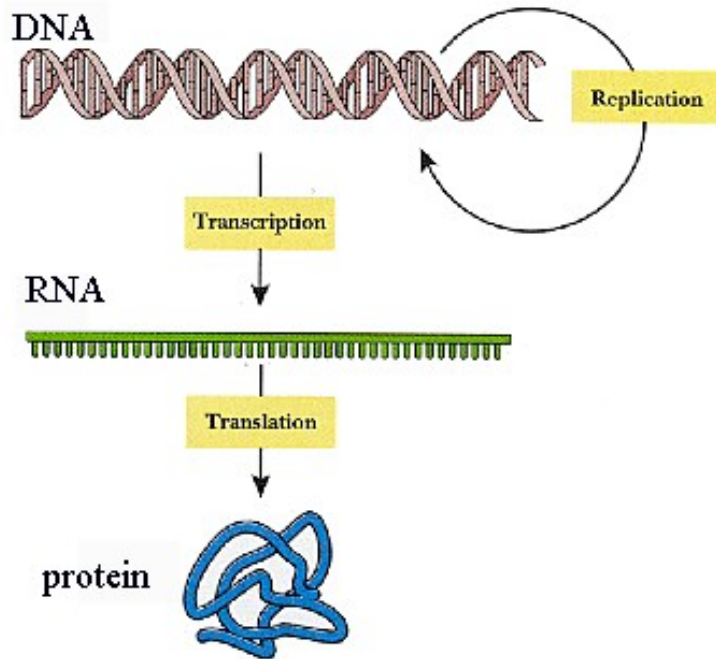
- DNA molecule is a linear polymer
- Structure can be represented as string of 4 symbols: ACTG
- These “sequences” can be analyzed mathematically/linguistically

HIV genome (approximately 10.000 nucleotides)

CGCAAGCGAAAGTAGAGCCAGAGGAGATCTCTCGACGCGAGGACTCGGCTTGCTGAAGTGCACCTCGGCAAGAGGCGAGAGCGGCGACTGGTGAGTACGCCATTTATATTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCCGAGAGCGTCAATATTAAG
AGGCAGAAAATAGTAAATGGGAGAAAATTAGTTTAAGCCAGGGGGAAGAACACATCTATGCTAAGGCAAGCAGGAGGCTGGAAGAGTTTGCACTTAACCTGGCCCTTTAGAAACAGCAGATGCTGTAAACAAATTAATAAACCA
GCTACAAACAGCTCTTAAGACGAGAAACAGGAGAACTTAGATCTATTGACACAGTAGCAACTCTCTATTGTGTACATAAAGGATAGATGTACGAGACACCAAGAGGCTTTAGACAAGATAGAGGAGAAACAAACAAAGTTTCAGCAAAAACACAGCA
GGCAAGGAGGCTGACGGGAAGGTCAGTCAAAATTTTCTTATAGTGCAGAATCTCCAAGGGCAAAATGGTACACCAGGCCATATCACCTAGAACTTTAAATGCATGGGTAAAGTGGTAGAAGAGAAGGCTTTTAGTCCAGAAGTAATACCCATGTTTTTCAGC
ATTATCAGAAGGAGCCACCCCAAGATTTTAAACACCATGCTAAACACAGTGGGGGACATCAAGCAGCCATGCATAATTTAAAGATACCATCAATGAAGAGGCTGCAGAATGGGATAGATTACATCCAGTACATGCAGGGCCATATCCAGTAGGCCAAAT
GAGAGAACCAAGGGGAAGTGACATAGCAGGAACACTAGTACCTTTCAGGAGCAAAATAGCATGGATGACAAGTAACCCACTGTTCCAGTAGGAGACATCTATAAAGATGGATAATTTCTGGGATTAATAAATAGTAAGAATTGTATAGCCCCACGACAT
TCTGGACATAAAACAGGGCCAAAGAACCTTTAGACCTGTAGACCGGTTCTTAAACTTTAAGAGCGGGAACACAGTACACAGATGTA AAAAATTTGGATGCAGACACACTTTGTGGTCCAAATTTGCCAACCAGATGTGAAGCAACTTCTAAGAGC
ATTAGGACAGGGGCTTCAATAGAAAGAAATGATGACAGCATGTGAGGAGTGGGAGGACCTAGTCATAAAGCAAGAGTGTGGCTGAGGCAATGAGCCAAACACAAATACCATAATGATGCAGAGAAGCAATTTTAAAGGCCCTAAAGAAATTTGTTAAATG
TTTCAACTGTGGCAAGAAAGGGCAGATAGCCAGAAATTTGAGGGCCCTAGGAAAAAGGCTGTTGGAAATGTGGAAGAAGGACACCAACTGAAAGATTGTACTGAGAGACAGGCTAATTTTTTAGGGAATACTGCGCCCTCCCAAGGGAAGGCCAGG
GAATTTTCTTCAGAGCAGACCAGGCCAACAGCCCCACAGAGGAGAGCTTCAGGCTTGGGGGAGAGACCAACTCCAGCTCAGAAGCAGGAGTCAACAGACAAGGAACATATCTCTTAACTCCCTCAGATCACTCTTTGGCAACGCCCTCGTCACA
ATAAAGATAGGGGGCAATTTAAAGGAAGCTCTATTAGATACAGGAGCAGATGATACAGTATTAGAAGACATGAATTTGCCAGGGAATGGAAACCAAAATGATAGGGGGAATGGAGGTTTTATCAAAGTAAGACAGTATGAACAAGTACCCATAGAAATC
TGTGGACACAAGCTATGGGTACAGATTAGTGGGACCTACACCTGTCAACATAATTTGGGAGAAATCTGTTGACTCAGCTTGGTTGCACTTTAAATTTTCCAATTAGTCCCATTTGAACTGTACACCATGTAAGATTAAGCCAGGAATGGATGGCCCAAGGTT
AAACAATGGCCATTGACAGAGAAGAAAATAAAGCATTTAACAGAAATTTGTAATGAAATGGGAAAGGAAGAAAATTAACAAAATTTGGGCTGAAATCCATATAACACTCCAAATTTGGCATAAAAGGAAGCACTAAGTGAAGGAAATAGTA
GATTTACAGGGAACCTCAATAAAAGAACTCAAGACTTTTGGGA
TTCAACCATCTAGTATATAATTAATGAACACCCAGGATCAG
ATTGATAACTTGTATGTAGGATCTGACTTAGATAGATAGGGCA
AAATGGACAGTACAGCCTATAAACTGCCAGAAAGGAAG
ATAGTACCATACTGAAGAGCAGAATTAGAATTGGCAGA
TTCAAAATCTAAAGACAGGGAATATGCAAAATAGGAGC
ACATGGTGGACAGACTTTGGCAGACCTGGATTCCTGA
AAAGCAGGTATGTTACTGACAGAGGAAGAAAAGGTTGT
GCACAACCATAGTAAGATGAATCAGATGTAGTTAACCAAT
TTAGATGGAAATAGATAAAGCTCAAGAAAGAACATGAAGATA
TGTAGTCCAGGGATTTGGCAATTAGATTGTACCCATTTAGA
GTCAAAGTATACATACAGACAATGGTAGTAACCTCACAG
CAGGTAAAGAGATCAAGCTGAGCCTTAAAGACAGCAGTACA
ACAAAATTTCAAAAATTTCCGGTTTATTAACAGACAGCAG
GGAAAACAGATGGCAGGTGCTGATTGTGTGGCAGGTAGACA
ACACATCCCAGTAGGAGAGCTAAATTAGTATAA AAAACAT
GTATTATTTGATTTTGGTTCAGACTCTGCCATAAGAAAG
TCTGCTAGTATTCAGAAATTAGTAGAGGATAGATGGAACA
CAATATATCTATGAACATATTTGGGATACTTTGGGCGAGGT
ACTAGAGCCTTGGAAAGCATCCAGGAAGTCCAGCTTAAGACT
CAGTGAGGATCATCAAAATCCTATATCAAAGCAGTAAGTAG
AATAGACAGATTAATTTAGATAGAATAAGAGAAAGCAGAAAG
ATTATGGGCTACCTGTATGGAAGAGCAACCACTTTA
TTAATATGTGGAAAGAAATGTAATGTAATCAGATGCATGAAG
ATAATACCTACCTTGAGGGCATGGATAAGGAAATAAAAAATTTGCTCTTTTAAATACAACACAGCTAATTAAGAGATAGAAAGCAGACAGGGTATGCACATTTTTATAGACTTGATGTAGTACCATTTGGTGAGAGAACTCTAGTGGGAACCTCTAGTGGGTATT
ATACATTAATAAATTTGTAATCACTGACCCATAACACAGGCTCTCCAAAGGCTCTCTTTTGATATCCATTCCTACATTTGCACTCCAGCTGGGTTATGCACTTTAAAGTGTAATAAATAGACATTTAATGGGACAGGACCATGCAATAATTTGTAGCACAG
TCAACTATACACATGGGATTAAGCCAGTGGTATCACTCACTACTGTTAAATGGTAGGTCAGAGATAAAATAAATAGATTAGTCTGAAAATCTGCAGAACCAATGCAAAAACATAATAGTACATCTTAAACCAATCTGTAGAAATTTGTATGCAACAGC
CCAACAATAATACGAAAAAGTATAAGGATAGGACCAGGACAACATTTCTATGCAACAGGAGAAAATAAGGAGACATAAGACAAGCATATTGTATCATTAAATGGAAGTCAATGGAATGATACCTTTACAAGAGTAAGTAAAAAATAGCAGACACTTCC
CAATAAATAACATAATTAATTTATCTCTCAGGGGGGACCTAGAAAATTAACAACATAGCTTTAATGTAGAGGAGAATTTTTCTATTTGTAATACATCAAAATTTGTTAATAGTACATACATGTCTAATGGTACATCTGTTAATGGTACAGACAGTA
ATTTCAACCTCAAACTACACATCAATCCCTGTCAGGATAAGCAAAATTAATAATTTGGGAGGAGTGAAGCAGGATATGTCGCCCTCCATTCGAGAAACATAACATCAATATCAAGATATCAGAGATTTACTATTGGTACGTGATGGAGGAGGGAATG
AGACACAAAATGATACAGAGACATTTCAGACCTGGAGGAGGAGATATGAGGAGCAATTTGGAGAAGTGAATTATATAAATATAAAGTGGTAGAAATTAAGCCATTAGGAGTAGTCTCCACTGGAGCAAAAAGGAGAGCGGTGGAGAGAGAAAAAGAGCAGCGG
GACTAGGAGCTTTGCTCTTTGGGTTCTTGGGAGCAGCAGGAAGCAGTATGGGCGCGCGCTCAATAACGCTGACGGTACAGGCCAGCAAAATGCTGTCCGATATAGTGAACAGCAAAAGCAATTTGCTGAGAGCTATAGAGGCGCAACAGCATCTGTTGCAAC
TCACGGTCTGGGCGATTAAAGCAGCTCCAGACAGCATCTGGCTATAGAAAAGTAACTAAAAGAGCAACAGCTCTTAGGCTTTGGGCTGTCTCGGAAAACTCATCTGCACCACTGATCTGACCTTTGGAATCTGGAATCTGGAATCTCAAGAACAGTA
TTTGGAACTACACACTTGATGCAATGGGATAGGAAATTAGTAATTAACAGACATTAATATACAGCTTTGATTTAGAAATTCGCAAAACACAGGAAAGGAATGCAAAAGGATCTATTAGAATTGGACATTTGGAACAACTTTGGAATTTGGTTTAAACATAT
CAAATTTGGCTGTGGTACATAAAAAATTTTCAATAAATAGTAGGAGGCTGATAGGTTTAAAGAAATTTTTTGTGCTGCTTTCTATAGTAAATAGAGTTAGGCGGGATATCACTCTTTGTGCTTTTCAGATCCCTACCCAGAACCAGGGGACCTCGACAGGC
TCGGAAGAAATCGAAGAAGGTGGAGAGCAAGCAAGACAGATCCATTCGATTAGTGAACGGATTCTTAGCTTTGCTGCTGGGACGATCTCGGAACCTGTGCCTCTTCAGTACCCACCGCTTAGAGACTTCATATTAGTGGTAGCGAGATGGTGGAAAC
TTCTGGAACCGCAAACTTCTCAGGGGACATACAGAGGGGCTGGGAAGCCCTTAAATACCTGGGAAGCCCTTGTGCAGTACTGGGTCAGGAGTCAAAAAGAGTGCATTAGTGTCTGTGATACCATAGCAATAGCAGTAGCTGAAGGAACAGATGGATTATAG
AATTAGTACAAGGCTTTTTAGGGCATCGGCAACGTAACCTAGAAGAATAAGACAGGGCTTTGAAGCAGCTTTGCAATAAAATGGGGGCAAGTGGTCAAAACGTAGCATAGTTGGATGGCTGCTATAAGGGAAGAATGAGAAGAACTCAGCCAGCAGCA
GATAGGTTGGAAGCAGTAATCTCGAGCAGCATGGAGTGGGAGCAGTATCTCGAGACCTGGAAAGACATGGAGCAATCAAGATGCAAAATACAGCACTACTAATGAGGCTGTGGCTGGCTAGAGACATCAGGAGGAGGAGGTGGGTTTCCGACTCAGA
CCTCAGGTACCTTTAAACCAACTGACTTCAACGAGCGCTGTAGATCTTAGCTTTTAAAAAGAAAGGGGCAAGTGAAGGTTAATTACTCTAAGAAAAGACAGAGAGATCTTGATTGTGGGCTCTACACACAAAGGCTACTTCCCTGACTGGCA
AACTACACACAGGACCAGGATCAGATTTCCACTGACCTTTGGGTGGTCTTCAAGCTAGTACCAGTTGACCCAGGGAAGTGAAGGAGACCAACGAAGGAGAGACAACCTGCCTGCTACACCCTGTGTGCCAGCATGGAATTGGAGGATGAACACAGAGAA
GTCTTAAAGTGAAGTTTGACAGTCACTAGCAGCAGACATGCCCCGCGAAGTACATCCCGAGTTTACAAGACTGCTGACACAGAAAGGACCTTCCGCGGGGACTTTCCACTGGGGGCTTCTGGGAGGTGTGGTCTGGGCGGAGTGGGAGTGGTCA
ACCTCAGATGCTGCATATAAGCAGTGTCTTTTGCTGTACTGGTCTCTCTAGTTGGACCAGATCTGAGCTGGGAGCTCTCTGGCTATCTCGCGAACC

ENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

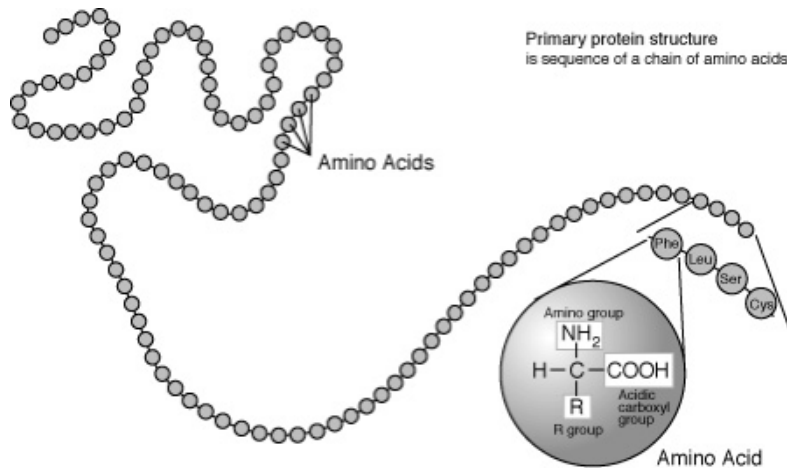
DNA --> RNA --> protein



Standard Genetic Code

	T			C			A			G			
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C	T
	TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C	C
	TTA	Leu	L	TCA	Ser	S	TAA	Och *		TGA	Opa *		A
	TTG	Leu	L	TCG	Ser	S	TAG	Amb *		TGG	Trp	W	G
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R	T
	CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R	C
	CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R	A
	CTG	Leu	L	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R	G
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S	T
	ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S	C
	ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R	A
	ATG	Met	M	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R	G
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G	T
	GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G	C
	GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G	A
	GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G	G

Symbolic representation of protein structure

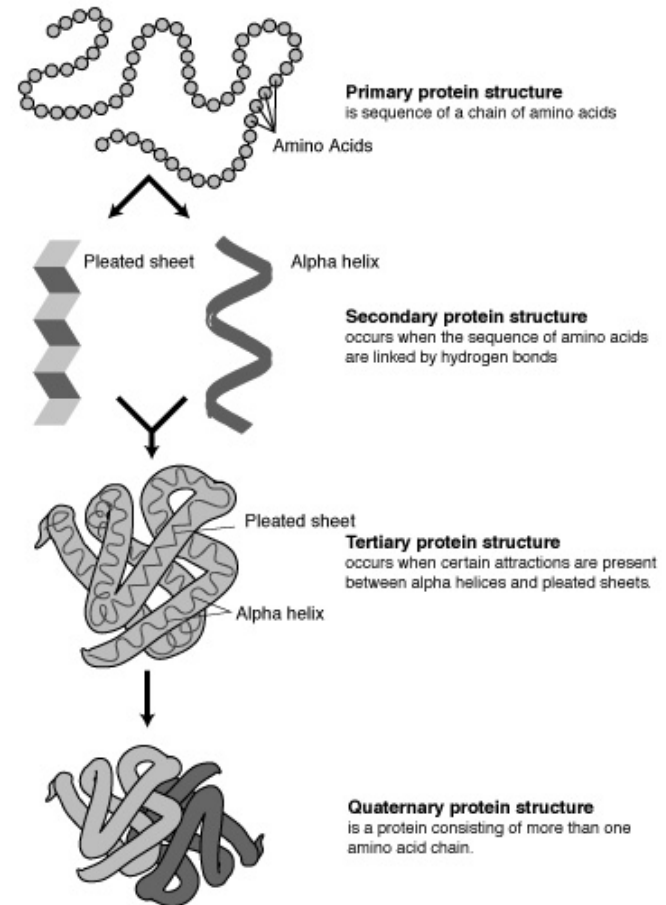


Proteins are linear polymers

Built from 20 amino acids

Can be represented as string of 20 symbols

ACDEFGHIKLMNPQRSTVWY



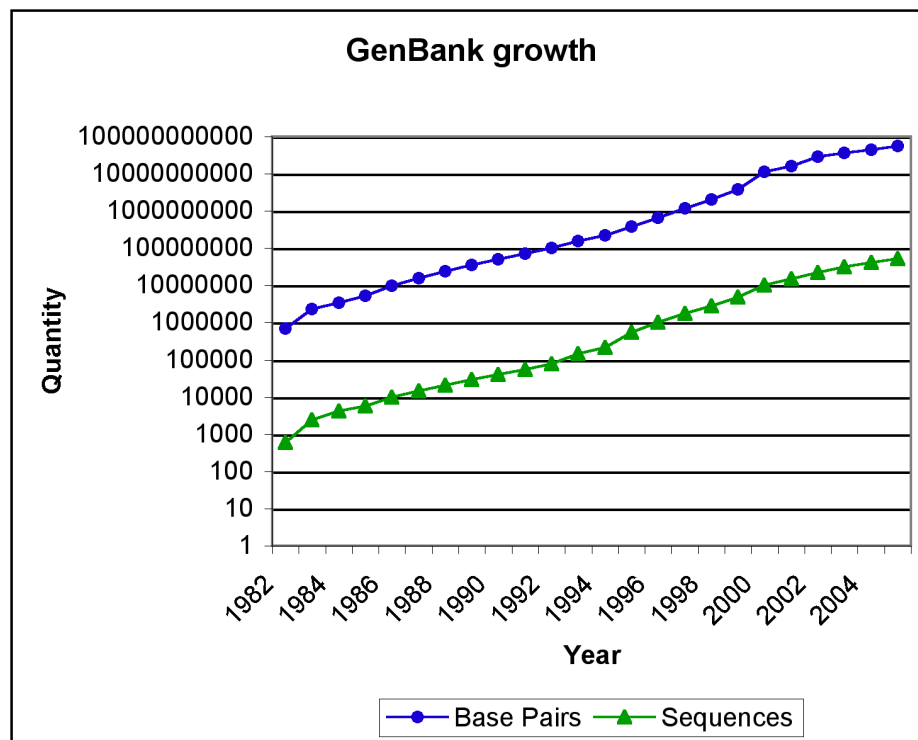
NCBI databases

The screenshot displays the NCBI (National Center for Biotechnology Information) website. At the top, the browser address bar shows the URL <http://www.ncbi.nlm.nih.gov/>. The page header includes the NCBI logo, navigation links like 'Resources' and 'How To', and a search bar with the text 'human globin' and a 'Search' button. The main content area is divided into several sections:

- Resources:** A vertical menu on the left lists various resources including 'NCBI Home', 'All Resources (A-Z)', 'Literature', 'DNA & RNA', 'Proteins', 'Sequence Analysis', 'Genes & Expression', 'Genomes', 'Maps & Markers', 'Domains & Structures', 'Genetics & Medicine', 'Taxonomy', 'Data & Software', 'Training & Tutorials', 'Homology', 'Small Molecules', and 'Variation'.
- Welcome to NCBI:** A central section with a welcome message and links to 'More about the NCBI', 'Mission', 'Organization', 'Research', and 'RSS'.
- Genome Reference Consortium:** A section featuring a graphic of DNA sequence and text describing the consortium's mission to improve human and mouse reference assemblies.
- How To...:** A section with a list of tasks such as 'Obtain the full text of an article', 'Retrieve all sequences for an organism or taxon', 'Find a homolog for a gene in another organism', 'Find genes associated with a phenotype or disease', 'Design PCR primers and check them for specificity', 'Find the function of a gene or gene product', and 'Determine conserved synteny between the genomes of two organisms'.
- Popular Resources:** A list of popular resources including 'PubMed', 'PubMed Central', 'Bookshelf', 'BLAST', 'Gene', 'Nucleotide', 'Protein', 'GEO', 'Conserved Domains', 'Structure', and 'PubChem'.
- NCBI News:** A section with news items dated '02 Dec 2009' and '05 Oct 2009', featuring links to 'November and October News' and 'NCBI News - September 2009'.
- NLM/NCBI H1N1 Flu Resources:** A section at the bottom with a link to 'See all ...'.

NCBI GenBank

- GenBank is one of the main international DNA databases.
- GenBank is hosted by NCBI: National Center for Biotechnology Information.
- GenBank has existed since 1982.
- The database is public - no restrictions on the use of the data within.



GenBank format

GenBank format

```
LOCUS       000000000      1185 bp    DNA     linear     VRT 18-APR-2005
DEFINITION  Cairina moschata (duck) gene for alpha-D globin.
ACCESSION   X01831
VERSION     X01831.1  GI:162724
KEYWORDS    alpha-globin; globin.
SOURCE      Cairina moschata (Muscovy duck)
  ORGANISM  Cairina moschata
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
REFERENCE   1  (bases 1 to 1185)
AUTHORS     Ertel, C. and Niesing, J.
TITLE       The primary structure of the duck alpha D-globin gene: an unusual
           5' splice junction sequence
JOURNAL     DMO 3, 2 (8), 1339-1343 (1983)
COMMENT     Data kindly reviewed (13-NOV-1985) by J. Niesing.
FEATURES             Location/Qualifiers
     source          1..1185
                     /organism="Cairina moschata"
                     /mol_type="genomic DNA"
                     /db_xref="tacon:8855"
     CDS             101..1114
                     /note="primary transcript"
                     /number=1
     repeat_region   227..246
                     /note="direct repeat 1"
                     /number=1
     repeat_region   289..309
                     /note="direct repeat 1"
                     /number=2
     repeat_region   387..391
                     /note="direct repeat 1"
                     /number=2
     repeat_region   592..599
                     /note="direct repeat 1"
                     /number=2
     repeat_region   940..1114
                     /note="direct repeat 1"
                     /number=3
     polyA_signal    1095..1100
     polyA_signal    1114
ORIGIN
1  ctggtgtggtc  taagcccttc  cacccttcca  cgtgtataag  ataaggccag  ggcgggagcg
41  cagggtgcta  taagagctcg  gcccgcgggg  tgtcttccac  acagaacacc  gtcagttgac
121  agctgtgac  cgcgtctgag  cactgtgtac  cgcgtgtgac  cgcgtgtgac  aagaagctca  tctgtgtggt
181  ctggagagag  ctggtgtggt  acagagagag  atctgagagt  gaagctgtcg  aaggtgtgtg
241  gctgtgtgca  ggggggactc  acaggtgtgg  cagcagggag  caggagccct  gcagcgggtg
301  tgggtgtgga  cccagagcgc  cagcgggtgt  ggtgtgagat  gggcaagcca  gcaggccacc
361  aaaaattgct  ggtgtgtgtc  cggcaggtgt  tctgtgtgtc  acccagagct  caagcttacc
421  ttcccccact  tgaactgtga  tccgtgtgtc  gaacaggttc  gtgtgtgtgt  caagaaggtg
481  ggggtgtgac  tggcgaatgc  cgtgagagag  ctggacacac  taagccagcg  cctgtctgag
541  ctacgaacac  tgaatgtgta  ccaactgtgt  gtgtgtgtgt  tcaactttaa  ggcagagggg
601  gactaggttc  cttgtgtgtg  ggggtgtgag  ggtgtgtgtg  gtcaggtgtg  ggggtccagg
661  ggttgtggtt  tctgtgtgtg  tggagttgtt  ggggtgtgag  ggcagaggtc  ctgtgtgtgt
721  ggttcaaggt  gttctgtgtg  cagcagagga  gacagaggtg  gttgtgtgtg  gttgtgtgtg
781  gtgggagaga  gctgtgtgtt  gtgtgtgtga  tggaggtgtg  gtcaggtgtg  ggcagaggtt
841  gggggactaa  ggtgtgtgtg  ggcaggtgtg  gggggactga  ggcagactga  ggcagactga
901  tccagagag  ggtgtgtgtg  cctgtgtgtg  cctgtgtgtg  gttgtgtgtg  gttgtgtgtg
961  tgtgtgtgtg  gtcagactgt  ggcagactgt  acagagagag  gtcagactgt  gttgtgtgtg
1021  agtgtgtgtg  gtcagactgt  gtcagactgt  gtcagactgt  gtcagactgt  gtcagactgt
1081  cctgtgtgtg  tcaactaaag  acacactaac  cagactgtgt  tctgtgtgtg  tctgtgtgtg
1141  ggggtgtgtg  gttgtgtgtg  aggtgtgtgt  tctgtgtgtg  cactgtgtgt  cactgtgtgt
```

Header

Indeholder information ang. Organisme, publikation, Accession ID mm.

FEATURE blok

Indeholder en beskrivelse af forskellige elementer i DNA sekvensen.

CDS: Coding Sequence. Indeholder koordinater på den protein kodende del af et gen. Bemærk de tre intervaller.

ORIGIN blok

Indeholder selve DNA sekvensen.

- Originates from the GenBank database.
- Contains both a DNA sequence and annotation of feature (e.g. Location of genes).

GenBank format - HEADER

LOCUS CMGLOAD 1185 bp DNA linear VRT 18-APR-2005
 DEFINITION Cairina moschata (duck) gene for alpha-D globin.
 ACCESSION X01831
 VERSION X01831.1 GI:62724
 KEYWORDS alpha-globin; globin.
 SOURCE Cairina moschata (Muscovy duck)
 ORGANISM Cairina moschata
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Archosauria; Aves; Neognathae; Anseriformes; Anatidae; Cairina.
 REFERENCE 1 (bases 1 to 1185)
 AUTHORS Erbil,C. and Niessing,J.
 TITLE The primary structure of the duck alpha D-globin gene: an unusual
 5' splice junction sequence
 JOURNAL EMBO J. 2 (8), 1339-1343 (1983)
 PUBMED 10872328
 COMMENT Data kindly reviewed (13-NOV-1985) by J. Niessing.

GenBank format - ORIGIN section

ORIGIN

```

1  ctgcgtggcc tcagcccctc caccctcca cgctgataag ataaggccag ggcgggagcg
61  cagggtgcta taagagctcg gcccgcggg tgtctccacc acagaaaccc gtcagttgcc
121 agcctgccac gccgctgccg ccatgctgac cgccgaggac aagaagctca tcgtgcaggt
181 gtgggagaag gtggctggcc accaggagga attcggaagt gaagctctgc agaggtgtgg
241 gctgggcca gggggcactc acagggtggg cagcaggag caggagccct gcagcgggtg
301 tgggctggga cccagagcgc cacggggtgc gggctgagat gggcaaagca gcagggcacc
361 aaaactgact ggcctcgctc cggcaggatg ttctctgcct acccccagac caagacctac
421 ttccccact tcgacctgca tcccggtctt gaacagggtc gtggccatgg caagaaagtg
481 gcggtgccc tgggcaatgc cgtgaagagc ctggacaacc tcagccaggc cctgtctgag
541 ctcagcaacc tgcattgcta caacctgcgt gttgacctg tcaacttcaa ggcaagcggg
601 gactagggtc cttgggtctg ggggtctgag ggtgtgggt gcagggtctg ggggtccagg
661 ggtctgagtt tcctggggtc tggcagtcct gggggctgag ggccagggtc ctgtggtctt
721 gggtagcagg gtcctggggg ccagcagcca gacagcaggg gctgggattg catctgggat
781 gtgggcccaga ggctgggatt gtgtttggaa tgggagctgg gcaggggcta gggccagggt
841 gggggactca gggcctcagg gggactcggg gggggactga gggagactca gggccatctg
901 tccggagcag gggactaag ccctggtttg ccttgagct gctggcacag tgcttccagg
961 tgggtgctggc cgcacacctg ggcaaagact acagccccga gatgcatgct gcctttgaca
1021 agttcttgct cgccgtggct gccgtgctgg ctgaaaagta cagatgagcc actgcctgca
1081 cccttgacc ttcaataaag acaccattac cacagctctg tgtctgtgtg tgctgggact
1141 gggcatcggg ggtcccaggg agggctgggt tgcttcaca catcc

```

//

GenBank format - FEATURE section

```

FEATURES                     Location/Qualifiers
    source                    1..1185
                              /organism="Cairina moschata"
                              /mol_type="genomic DNA"
                              /db_xref="taxon:8855"
    CAAT_signal               20..24
    TATA_signal               69..73
    precursor_RNA             101..1114
                              /note="primary transcript"
    exon                      101..234
                              /number=1
    CDS                       join(143..234,387..591,939..1067)
                              /codon_start=1
                              /product="alpha D-globin"
                              /protein_id="CAA25966.2"
                              /db_xref="GI:4455876"
                              /db_xref="GOA:P02003"
                              /db_xref="InterPro:IPR000971"
                              /db_xref="InterPro:IPR002338"
                              /db_xref="InterPro:IPR002340"
                              /db_xref="InterPro:IPR009050"
                              /db_xref="UniProt/Swiss-Prot:P02003"
                              /translation="MLTAEDKKLIVQVWEKVGHQEEFGSEALQRMFLAYPQTKTYFP
HFDLHPGSEQVRGHGKKVAAALGNVKS L DNL SQALSEL SNLHAYNLRVDPVNFKLLA
QCFQVVLA AHLGKDYSPEMHAAFDKFLSAVA AVLA EKYR"
    repeat_region             227..246
                              /note="direct repeat 1"
    intron                    235..386
                              /number=1
    repeat_region             289..309
                              /note="direct repeat 1"
    exon                      387..591
                              /number=2
    intron                    592..939
                              /number=2
    exon                      940..1114
                              /number=3
    polyA_signal              1095..1100
    polyA_signal              1114
  
```

NCBI databases: fasta format

The screenshot displays the NCBI Nucleotide database interface. The browser address bar shows the URL: [http://www.ncbi.nlm.nih.gov/nuccore/28302130?report=fasta&log\\$=seqview&from=54&to=497](http://www.ncbi.nlm.nih.gov/nuccore/28302130?report=fasta&log$=seqview&from=54&to=497). The page title is "Nucleotide - Homo sapiens hemoglobin, gamma A (HBG1), mRNA". The NCBI logo is visible in the top left. The search bar contains "Nucleotide" and "for". The format dropdown is set to "FASTA". The sequence is displayed in FASTA format, starting with >gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA. The sequence is: ATGGGTCATTTACAGAGGAGGACAAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG CTGGAGGAGAAACCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG CAACCTGTCTCTGCCTCTGCCATCATGGGCAACCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCACG GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTGGCAAT CCATTTGGGCAAGAATTACACCCCTGAGGTGCAGGCTTCTGGCAGAAGATGGTGACTGCAGTGGCCAGT GCCCTGTCTCCAGATACCACTGA. The right sidebar contains sections: "Change Region Shown" (Whole sequence, Selected Region), "Customize View", "Analyze This Sequence" (Run BLAST, Pick Primers), "Articles about the HBG1 gene" (Molecular analysis of gamma-globin promoters, A genome-wide association identified the common genetic variant, Expression of miR-210 during erythroid differentiation and induction), "RefSeq Protein Product" (See the reference protein sequence for A-gamma globin (NP_000550.2)), and "More about the HBG1 gene" (The gamma-globin genes (HBG1 and HBG2)).

Nucleotide - Homo sapiens hemoglobin, gamma A (HBG1), mRNA

http://www.ncbi.nlm.nih.gov/nuccore/28302130?report=fasta&log\$=seqview&from=54&to=497

NCBI

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Nucleotide for

Format: GenBank FASTA Graphics More Formats

Showing 444 bp region from base 54 to 497.

NCBI Reference Sequence: NM_000559.2

Homo sapiens hemoglobin, gamma A (HBG1), mRNA

>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTACAGAGGAGGACAAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGAGGAGAAACCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG
CAACCTGTCTCTGCCTCTGCCATCATGGGCAACCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT
TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCACG
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTGGCAAT
CCATTTGGGCAAGAATTACACCCCTGAGGTGCAGGCTTCTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCTCCAGATACCACTGA

Change Region Shown

☐ Whole sequence
☒ Selected Region
from: 54 to: 497
Update View

Customize View

Analyze This Sequence

- Run BLAST
- Pick Primers

Articles about the HBG1 gene

- Molecular analysis of gamma-globin promoters, HS-111 and [Hemoglobin. 2009]
- A genome-wide association identified the common genetic variant [Hum Genet. 2009]
- Expression of miR-210 during erythroid differentiation and induction [BMB Rep. 2009]

» See all...

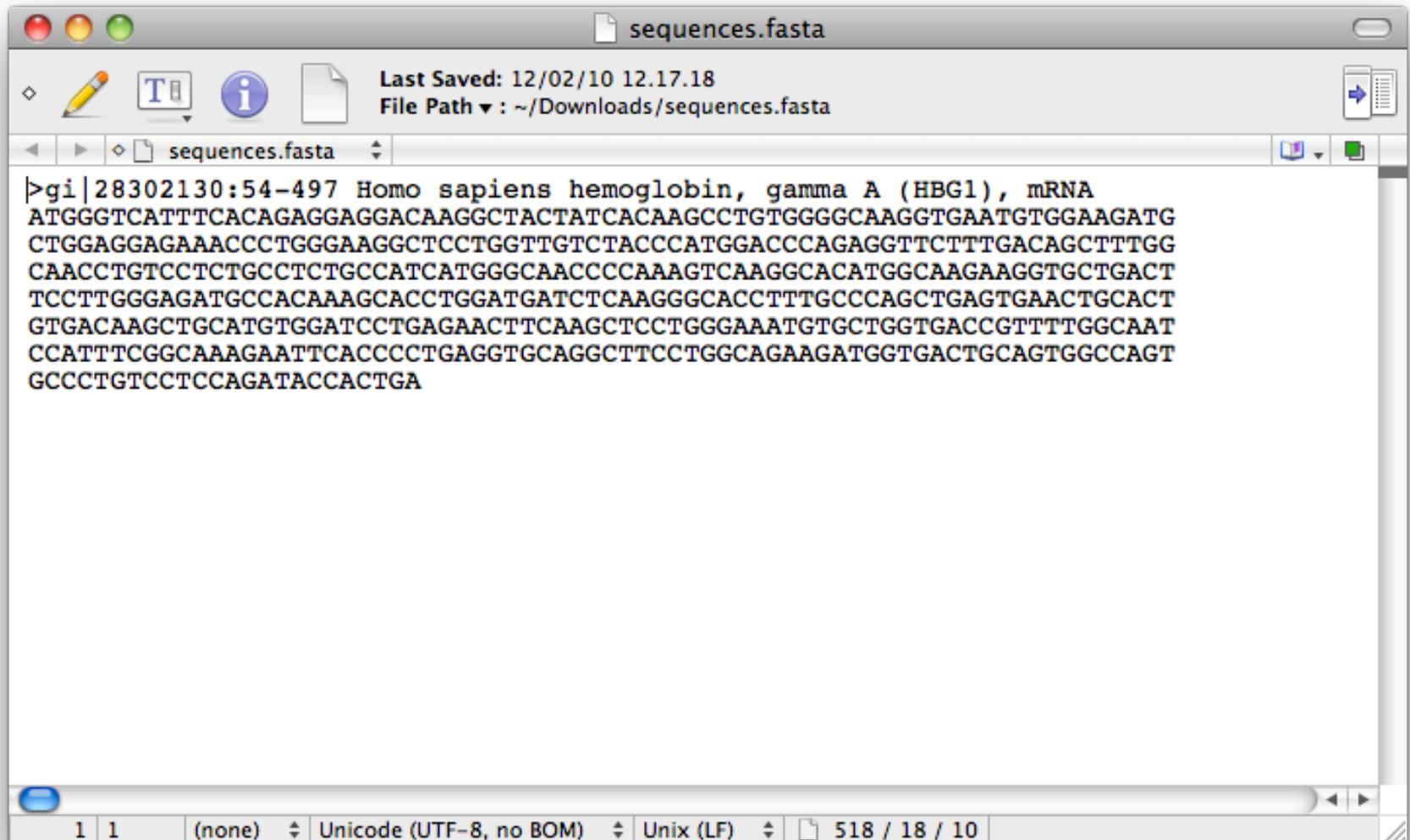
RefSeq Protein Product

See the reference protein sequence for A-gamma globin (NP_000550.2).

More about the HBG1 gene

The gamma-globin genes (HBG1 and HBG2)

FASTA file



The image shows a screenshot of a text editor window titled "sequences.fasta". The window has a standard macOS-style title bar with red, yellow, and green buttons. Below the title bar is a toolbar with icons for undo, redo, save, and other functions. The main text area contains a FASTA file entry for "Homo sapiens hemoglobin, gamma A (HBG1), mRNA". The sequence is displayed in a monospaced font, with the header line starting with ">gi|28302130:54-497" and the sequence lines following. The status bar at the bottom shows the current line and column (1 / 1), the encoding (Unicode (UTF-8, no BOM)), the line ending (Unix (LF)), and the total file size (518 / 18 / 10).

```
>gi|28302130:54-497 Homo sapiens hemoglobin, gamma A (HBG1), mRNA
ATGGGTCATTTACAGAGGAGGACAAGGCTACTATCACAAGCCTGTGGGGCAAGGTGAATGTGGAAGATG
CTGGAGGAGAAACCCTGGGAAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTTTGACAGCTTTGG
CAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGCTGACT
TCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCAGCTGAGTGAAGTGCAGT
GTGACAAGCTGCATGTGGATCCTGAGAACTTCAAGCTCCTGGGAAATGTGCTGGTGACCGTTTTGGCAAT
CCATTTTCGGCAAAGAATTCACCCCTGAGGTGCAGGCTTCTGGCAGAAGATGGTGACTGCAGTGGCCAGT
GCCCTGTCCTCCAGATACCACTGA
```